RULEROUTER: SPARSE, TRACEABLE NATURAL-LANGUAGE INTERVENTIONS FOR FROZEN VISION-LANGUAGE MODELS

Anonymous authors
Paper under double-blind review

ABSTRACT

Natural-language rules can target structured errors in frozen vision-language models, yet unconditional application often overturns correct predictions. We introduce RuleRouter, a sparse and traceable intervention layer that derives contrastive rules from development-only confusion evidence and applies a proposed correction only when at least four of five benefit–harm router replicas agree that predicted benefit exceeds harm. The rules, routers, and operating points are frozen before their corresponding confirmation evaluations. Relative to a supervised logit-adapted System-1, RuleRouter improves Macro-F1 on EuroSAT-RGB by 5.53 percentage points (95% paired-bootstrap CI $[3.90, 7.06]$) at 16.8% intervention coverage. On an independent AID-30 final-confirm set, a separately frozen pipeline improves its prespecified reporting comparator by 3.17 points (CI $[2.26, 4.15]$), while changing 7.83% of its predictions. So2Sat serves as a low-base-performance stress test and shows a smaller 0.46-point gain. Across these evaluations, selective routing produces more corrections than harms and avoids the degradation caused by unconditional rule fusion. Equal-label visual probes remain more accurate overall, while matched semantic controls favor confusion-driven rules on EuroSAT but provide no corresponding evidence on So2Sat. RuleRouter therefore addresses settings that require sparse changes to a frozen predictor, explicit coverage and harm accounting, and an inspectable textual record for each intervention.

1 INTRODUCTION

Frozen vision-language models (VLMs) classify images through text prompts without task-specific weight updates (Radford et al., 2021; Bai et al., 2025). We call their rule-free predictions System-1 (S1). Repeated confusions between related classes suggest targeted natural-language rules. Yet the same rule can repair one prediction and overwrite another that was already correct (Figure 1).

Confidence in the current prediction does not tell us whether a proposed correction will help. In the frozen V2 comparison, unconditional rule fusion trails the selective policy by 11.38 Macro-F1 points on EuroSAT and 4.50 points on the So2Sat stress test. The decision is therefore not just which rule to propose, but when to apply it.

RuleRouter fits a supervised logit adapter, constructs contrastive rules from development confusions, and uses five benefit–harm router replicas to gate each proposal. Intervention requires positive estimated utility and 4-of-5 agreement; every activated decision has a textual rule trace. Otherwise, RuleRouter retains adapted S1.

We evaluate this decision against prespecified comparators on three confirmation sets. The EuroSAT and independent AID pipelines gain $5.53\,pp$ and $3.17\,pp$ while changing 16.8% and 7.83% of outputs, respectively; So2Sat provides a low-base-performance stress test. Unconditional-fusion, semantic, matched-coverage, and equal-label visual controls distinguish the benefit of selective intervention from broader claims of accuracy superiority.

Reviewers: please read the Reviewer Guidelines (iclr.cc/Conferences/2027/ReviewerGuidelines) and the AI Policy for Reviewers (iclr.cc/Conferences/2027/AIPolicyForReviewers). If you used AI to expand, edit, or polish your review, please provide the input text to the LLM. Better still, consider skipping the LLM and submitting your original text: we, and the authors, are much more interested in your unedited thoughts than in what an LLM has to say. AI-assisted or not, you are putting your name and reputation behind your review: LLM-generated falsehoods, hallucinations or misrepresentations are subject to disciplinary action, which may include desk-rejecting all papers you have authored.

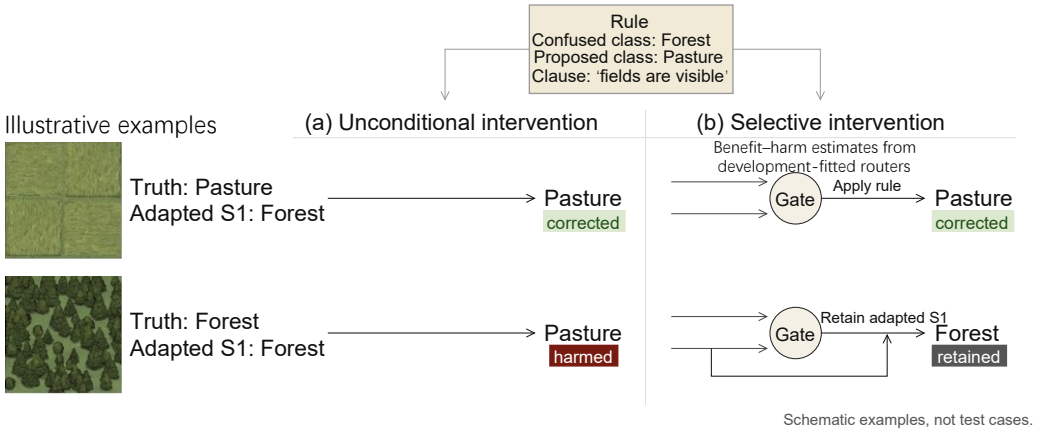


Figure 1: **Why selective intervention is needed.** The same Forest-to-Pasture proposal can correct an error or overwrite a correct prediction. (a) Unconditional intervention applies the proposal in both cases. (b) A development-fitted benefit–harm router decides whether to intervene or retain adapted S1. These are schematic examples, not test cases; ground-truth labels illustrate the outcomes and are not gate inputs.

2 Related Work

2.1 Vision-Language Transfer and Language-Based Concepts

CLIP-style models align images and text for zero-shot recognition (Radford et al., 2021); Qwen2.5-VL extends multimodal instruction following (Bai et al., 2025), and RemoteCLIP adapts pretraining to remote-sensing imagery (Liu et al., 2024). Concept bottlenecks and language-derived concept methods expose intermediate descriptions (Koh et al., 2020; Yang et al., 2023; Menon & Vondrick, 2023; Oikarinen et al., 2023). RuleRouter keeps the VLM weights frozen and uses development-confusion-driven class-pair rules as optional corrections to an adapted prediction.

Concept bottleneck models predict through concepts, while classification by description and label-free concept bottlenecks organize recognition around textual attributes (Koh et al., 2020; Menon & Vondrick, 2023; Oikarinen et al., 2023). Our class-pair rules instead act on selected predictions. Each intervention records the triggering clause and proposed label; the record does not explain how the VLM internally represented the image.

2.2 Selective Prediction, Deferral, and Risk-Aware Intervention

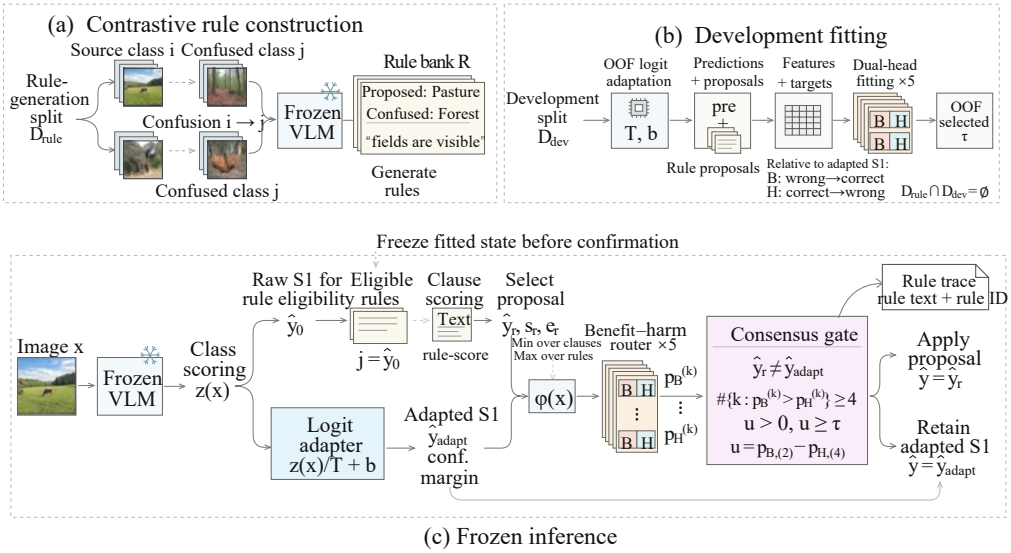
Temperature scaling adjusts confidence (Guo et al., 2017); selective prediction can reject uncertain outputs (Geifman & El-Yaniv, 2019), and learning-to-defer can hand a case to another decision maker (Mozannar & Sontag, 2020). Our action instead replaces a specific label with a rule-proposed label, so both corrected and harmed outputs can be counted against adapted S1. A readable rule trace does not establish that the VLM's internal reasoning followed the rule (Jacovi & Goldberg, 2020). We report empirical benefit and harm rather than a conformal risk guarantee (Angelopoulos et al., 2024).

Rejection decides whether to deliver a prediction; deferral decides who makes it. Rule intervention instead changes a delivered label while leaving VLM weights unchanged. The router therefore compares two actions: retaining the adapted prediction or applying the proposed rule. Development labels identify corrections and new errors for this comparison, and confirmation data supply their counts at a fixed operating point. These counts do not bound risk on future samples.

3 Method

Figure 2 follows RuleRouter from rule construction through development fitting to frozen inference. Raw S1 determines rule eligibility; adapted S1 defines both the benefit–harm comparison and the

fallback. The EuroSAT/So2Sat route is described below. Section 3.5 distinguishes AID's separately frozen fallback from its reporting comparator.



(c) Frozen inference

Figure 2: **RuleRouter overview.** (a) A separate rule-generation split supplies class-pair confusion evidence for contrastive rules. (b) Development data fit the logit adapter and five bootstrap benefit–harm routers, and select the out-of-fold cutoff τ. (c) Frozen inference uses raw S1 for rule eligibility and adapted S1 for the intervention comparison and fallback. A changed proposal is accepted only with at least four benefit-over-harm votes, $u > 0$, and $u \geq \tau$; accepted interventions retain a textual rule trace.

3.1 Framework Overview

RuleRouter separates contrastive rule construction, development-time fitting, and frozen inference so that rule proposals can be evaluated before they are allowed to modify the base prediction.

We first construct contrastive rules on a separate rule-generation split. For each directed confusion in which S1 predicts class j for examples from source class i, the frozen VLM compares examples from the two classes and generates short visual clauses that favor i over j. These clauses form a fixed bank of candidate corrections. Whether a candidate is applied is determined later by the routing mechanism.

Development data are then used to fit both the supervised logit adaptation and the routing policy. The temperature and class-bias parameters define adapted S1. For each rule proposal, out-of-fold predictions record separately whether the proposal repairs an adapted-S1 error or overwrites a correct adapted-S1 prediction. Five bootstrap dual-head routers learn these benefit and harm outcomes from prediction and rule-evidence features. The operating cutoff is selected using the same out-of-fold development procedure, after which the fitted components are frozen for confirmation.

At inference time, raw S1 determines the eligible rules for a new image, and clause scores select a single proposal. The five frozen routers estimate its benefit and harm relative to adapted S1. The proposal replaces the adapted label only when the replicas provide sufficient agreement and the utility cutoff is satisfied; otherwise, the prediction remains adapted S1. Each accepted intervention retains the triggering rule text and identifier. No ground-truth label is used in this decision.

Formally, let a frozen VLM produce a logit vector $\boldsymbol{z}(x) \in \mathbb{R}^C$ for an input image x, where the native label space is $\mathcal{Y} = \{1, \ldots, C\}$. Its direct prediction is

$$\hat{y}_0(x) = \arg\max_c z_c(x). \tag{1}$$

We refer to $\hat{y}_0(x)$ as the raw S1 prediction. A supervised post-hoc transform produces $\hat{y}_{\text{adapt}}(x)$, denoted by adapted S1. Using the raw S1 prediction to identify eligible contrastive rules, the rule module returns an alternative label $\hat{y}_r(x)$ together with its proposal strength $s_r(x)$.

The final prediction switches between the proposed rule correction and adapted S1 according to a binary routing decision $a(x) \in \{0, 1\}$:

$$\hat{y}(x) = \begin{cases} \hat{y}_r(x), & a(x) = 1, \\ \hat{y}_{\text{adapt}}(x), & a(x) = 0. \end{cases} \tag{2}$$

Raw S1 identifies the confusion edge for which a rule was constructed, whereas adapted S1 is the prediction that an accepted rule would replace. Rule eligibility therefore identifies a candidate correction but does not determine whether applying it is beneficial. The router makes this decision by estimating the relative utility of replacing the adapted prediction rather than retaining it.

3.2 Supervised System-1 Adaptation

We fit a temperature $T > 0$ and zero-sum class-bias vector $\boldsymbol{b} \in \mathbb{R}^C$ to development logits by minimizing cross-entropy. The VLM parameters remain frozen. The resulting distribution is

$$p_{\text{adapt}}(y = c \mid x) = \left[\text{softmax}\left(\frac{\boldsymbol{z}(x)}{T} + \boldsymbol{b}\right)\right]_c, \qquad \sum_{c=1}^{C} b_c = 0. \tag{3}$$

The corresponding prediction is

$$\hat{y}_{\text{adapt}}(x) = \arg\max_c p_{\text{adapt}}(y = c \mid x). \tag{4}$$

Positive temperature scaling preserves logit ordering, whereas the learned class bias can change the argmax. We therefore call the combined transform supervised post-hoc adaptation, not calibration alone.

We estimate the transform using nested out-of-fold development predictions and freeze it before confirmation. adapted S1 then serves as the prespecified comparator for router targets and intervention utility. [1]

3.3 Confusion-Driven Contrastive Rule Proposal

For EuroSAT and So2Sat, we construct directed confusion edges $i \to j$ on a dedicated rule-generation split that is disjoint from router-development and confirmation data. For each edge, the frozen VLM is prompted with three representative images from each class to generate visual evidence favoring source class i over confused class j. Each accepted rule contains at most two clauses and is filtered by class-name exclusion and syntactic checks. Full prompts and the rule catalog are provided in Appendices K and L.

At inference time, a rule $i \to j$ is eligible when the raw System-1 prediction is class j, and it proposes class i as an alternative. Rule strength is the minimum score across its clauses. Among eligible rules, we select the proposal with maximum strength, using source-class and rule IDs for deterministic tie-breaking. This yields a proposed label $\hat{y}_r(x)$ and strength $s_r(x)$.

Because logit adaptation can change the S1 label, a proposal selected from raw S1 may not improve on adapted S1. Benefit and harm are therefore defined against adapted S1, using development data separate from rule generation.

3.4 Utility-Aware Benefit–Harm Routing

The router estimates whether replacing adapted S1 with a rule proposal would correct an error or create one. For an example with $\hat{y}_r(x) \neq \hat{y}_{\text{adapt}}(x)$ and ground-truth label y, the benefit and harm

[1] For compatibility with archived experimental artifacts, the same transform is referred to as "calibrated S1" in some protocol files.

targets are defined as

$$y_B = \mathbb{1}\left[\hat{y}_r = y \wedge \hat{y}_{\text{adapt}} \neq y\right], \tag{5}$$

$$y_H = \mathbb{1}\left[\hat{y}_r \neq y \wedge \hat{y}_{\text{adapt}} = y\right]. \tag{6}$$

The two targets separate corrections (y_B) from new errors (y_H).

The router uses adapted confidence and margin, selected-rule strength, class-level rule evidence, the adapted and proposed classes, and whether the proposal changes the label. Let $\boldsymbol{e}_r(x) \in \mathbb{R}^C$ denote the class-rule evidence vector. We write the router input as

$$\begin{aligned}\varphi(x) = \big[&p_{\max}(x),\, m(x),\, s_r(x),\, \boldsymbol{e}_r(x),\\ &\text{onehot}(\hat{y}_{\text{adapt}}),\, \text{onehot}(\hat{y}_r),\, \mathbb{1}[\hat{y}_r \neq \hat{y}_{\text{adapt}}]\big],\end{aligned} \tag{7}$$

which has dimension $3C + 4$. Separate L1-regularized logistic heads are then fitted to estimate

$$p_B(x) = P(y_B = 1 \mid \varphi(x)), \qquad p_H(x) = P(y_H = 1 \mid \varphi(x)). \tag{8}$$

We train five bootstrap router replicas. For replica k, let $p_B^{(k)}(x)$ and $p_H^{(k)}(x)$ denote the corresponding benefit and harm estimates. A proposal requires at least four estimates to favor benefit over harm.

We further use a conservative margin across replicas,

$$u(x) = p_{B,(2)}(x) - p_{H,(4)}(x), \tag{9}$$

where $p_{B,(2)}$ is the second-smallest benefit estimate and $p_{H,(4)}$ is the second-largest harm estimate among the five replicas. This margin tolerates one dissenting replica without letting a single optimistic estimate determine the decision.

A rule is activated only when it changes the adapted prediction, receives benefit-over-harm support from at least four replicas, and satisfies

$$u(x) > 0, \qquad u(x) \geq \tau. \tag{10}$$

Otherwise, the system retains adapted S1.

3.5 Coverage Control and Fallback

The threshold τ sets intervention coverage. For EuroSAT and So2Sat, we select it from out-of-fold development predictions over the predeclared activation-coverage grid

$$\mathcal{G} = \{0,\, 1,\, 2.5,\, 5,\, 7.5,\, 10,\, 15\}\%. \tag{11}$$

The routers and selected threshold are frozen before confirmation; test labels do not enter their selection.

AID follows its separately prespecified gate budget of 156 images, equivalent to 6.5% activation coverage. For the EuroSAT and So2Sat route, abstention directly returns the reporting comparator adapted S1. AID preserves a separately frozen internal fallback that is not identical to its reporting System-1 comparator. Consequently, gate activation and changes relative to the reported comparator are treated as distinct quantities for AID. The complete accounting procedure is provided in Appendix E.

4 Experiments

4.1 Experimental Setup

We evaluate RuleRouter on the native label spaces of several remote-sensing benchmarks without introducing cross-dataset class mappings. EuroSAT-RGB (10 classes) (Helber et al., 2019) and the RGB rendering of So2Sat LCZ42 (17 classes) (Zhu et al., 2020) provide the original confirmation sets. A later AID-30 final-confirm set contains 2,400 previously unopened images (Xia et al., 2017) and is evaluated with the separately frozen AID pipeline. Earlier AID, PatternNet-38 (Zhou et al.,

2018), and NWPU-RESISC45 (Cheng et al., 2017) are retained as historical evidence. UC Merced-21 (Yang & Newsam, 2010) is retrospective and is therefore excluded from averages over independent confirmation results.

The experimental protocol separates rule construction from model fitting and final evaluation. For EuroSAT and So2Sat, each native class is partitioned into disjoint rule-generation, development, and sealed confirmation subsets. The rule-generation partition is used only to identify confusion pairs and construct contrastive rules, whereas development labels are used to fit the temperature and class bias, train the five router replicas, and select the operating point. File identities, content hashes, perceptual duplicates, and partition boundaries are checked before confirmation scoring. Once the development procedure is complete, the rules, logit adapter, router replicas, and operating thresholds remain fixed.

All confirmation routes use frozen Qwen2.5-VL-7B-Instruct (Bai et al., 2025); the VLM itself is never fine-tuned for the target datasets. The router replicas use the same fixed feature construction and are trained only from development evidence. This separation is important because the confirmation labels play no role in rule generation, router fitting, or operating-point selection. The second-backbone analysis is restricted to development data and is not included as independent confirmation evidence.

Our primary metric is image-level Macro-F1. EuroSAT and So2Sat are compared against cross-fitted adapted S1, while AID retains its prespecified frozen S1 reporting comparator. In addition to the full RuleRouter pipeline, we examine unconditional rule fusion, confidence-based and selective alternatives, visual baselines, router ablations, and matched semantic controls. We report activation coverage together with corrected and harmed predictions so that an accuracy gain can be interpreted alongside the amount and consequence of intervention. Statistical uncertainty is summarized using 95% paired stratified-bootstrap intervals with 10,000 replicates. Full partition sizes, fitting details, and inference settings are provided in Appendix A.

4.2 Main Results and Analysis

The frozen RuleRouter pipeline improves its prespecified System-1 comparator on all three independent confirmation sets, while modifying only a minority of predictions. On EuroSAT, Macro-F1 increases from 0.3168 to 0.3721, corresponding to a gain of $5.527\,pp$ (95% CI $[3.903, 7.061]$; Table 1, Figure 3). The router activates on 168 of the 1,000 confirmation images, or 16.8% of the set. Among these interventions, 78 predictions are corrected and 16 are harmed. Thus, the observed improvement is obtained without broadly replacing the adapted-S1 predictions: most inputs retain the frozen comparator, while the activated subset contains substantially more corrections than newly introduced errors.

AID final-confirm provides a second independent evaluation under a separately frozen protocol. Macro-F1 improves from 0.6915 to 0.7232 ($+3.172\,pp$, 95% CI $[2.256, 4.148]$). Relative to the reporting S1, 188 of the 2,400 outputs change, corresponding to 7.83% of the evaluation set. These changes include 104 corrections, 27 harms, and 57 wrong-to-wrong transitions. The intervention gate itself activates on 156 images, of which 151 alter the reporting S1. A further 37 changes arise because the frozen internal AID fallback is not identical to the reporting comparator. Gate activity and comparator-relative changes are therefore reported separately: the Macro-F1 gain measures the behavior of the complete frozen AID pipeline rather than the intervention gate in isolation (Appendix E).

So2Sat yields a smaller gain of $+0.456\,pp$ (95% CI $[0.167, 0.770]$), increasing Macro-F1 from 0.0597 to 0.0643. The router activates on 149 images, producing 17 corrections and no observed harms. Although the paired gain is positive, the very low adapted-S1 performance changes its interpretation. We therefore retain So2Sat as a low-base-performance stress test rather than treating it as evidence of practically strong scene recognition.

Taken together, the three confirmation results show a consistent direction of paired improvement but also illustrate why performance gain alone is insufficient to characterize the intervention. EuroSAT and AID combine positive Macro-F1 changes with limited comparator modification and explicit correction–harm accounting, whereas So2Sat exposes the behavior of the same selective mechanism when the underlying predictor is weak. The following analyses therefore examine not only whether routing improves the comparator, but also which components are responsible for the gain and how the result changes under stronger selective and visual baselines.

Table 1: **Independent confirmation relative to the named S1 comparator.** Gain is Macro-F1 in percentage points, with a 95% paired stratified-bootstrap interval. Gate denotes routed images; C/H counts all corrections/harms relative to the reporting comparator.

Dataset	S1	RuleRouter	Gain [95% CI]	Gate	C/H
EuroSAT	0.3168	0.3721	+5.53 $[+3.90, +7.06]$	16.80%	78/16
AID final-confirm	0.6915	0.7232	+3.17 $[+2.26, +4.15]$	6.50%	104/27
So2Sat (stress)	0.0597	0.0643	+0.46 $[+0.17, +0.77]$	8.76%	17/0

EuroSAT/So2Sat use adapted S1 ($n = 1{,}000/1{,}700$); AID uses its prespecified reporting S1 ($n = 2{,}400$). AID has 6.5% gate activity but 7.83% total changes relative to that comparator; Appendix E reconciles the two.

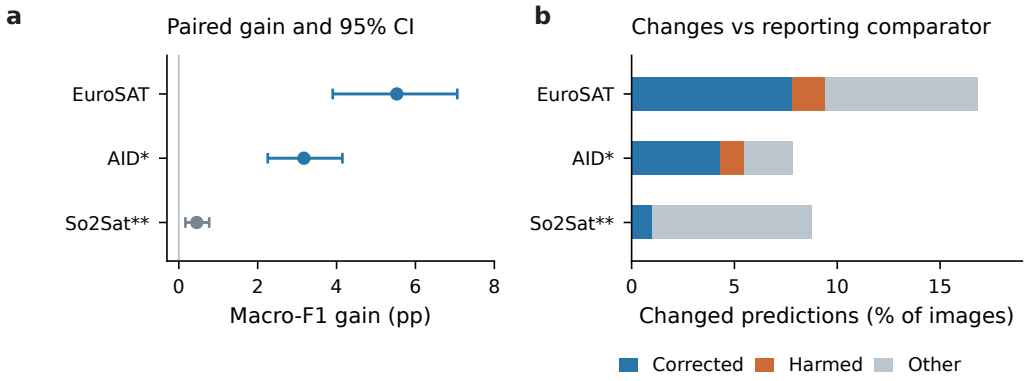


Figure 3: **Independent confirmation against the named comparator.** (a) Macro-F1 gains with 95% paired stratified-bootstrap intervals (10,000 replicates). (b) Corrected, harmed, and other changed predictions as percentages of all images. Other changes are wrong-to-wrong predictions. AID* includes changes outside its gate; So2Sat** is a stress test.

4.3 Historical multi-dataset results and the stronger-baseline boundary

The historical V1 route improves on raw S1 across four datasets (Table 2; Figure 4). UC Merced is retrospective and excluded from independent averages. On the three independent historical tests, adapted S1 exceeds V1 by 6.766, 9.614, and 7.426 points on AID, PatternNet, and NWPU. V1 therefore establishes gains over raw S1, not the stronger adapted comparator, and is not pooled with the later confirmation sets.

As coverage increases, the historical gain–harm trade-off changes (Appendix Figure 8). Figure 9 shows a larger benefit for selectivity than for any individual router component, whose effects vary by dataset. These V1 sweeps describe the earlier policies; the V2 cutoff was selected separately.

Table 2: **Historical V1 evidence relative to raw S1.** UC Merced is retrospective and excluded from independent averages.

Dataset	Evidence tier	n	Raw S1	V1 routed	Gain (pp; 95% CI)	Coverage; corrected/harmed
AID-30	independent	2,400	0.6728	0.7146	+4.180 $[3.403, 4.999]$	6.50%; 104/4
PatternNet-38	independent	3,800	0.5915	0.6611	+6.961 $[6.221, 7.736]$	14.71%; 259/16
NWPU-45	independent	4,500	0.6213	0.7120	+9.071 $[8.206, 9.973]$	18.76%; 435/115
UC Merced-21	retrospective	1,827	0.7164	0.7639	+4.742 $[3.649, 5.816]$	8.70%; 99/32

4.4 Current V2 controls: selectivity helps, component gains are limited

The V2 controls use adapted S1 as the comparator. Unconditional fusion reaches 0.2584/0.0193 Macro-F1 on EuroSAT/So2Sat; the full policy exceeds it by $11.38\,pp$ (95% paired CI $[8.20, 14.66]$) and $4.50\,pp$ ($[3.37, 5.65]$). At shared development targets of 15%/10%, the full-versus-benefit-only differences are $+0.103/+0.050\,pp$, with both intervals crossing zero. Replica averaging has no

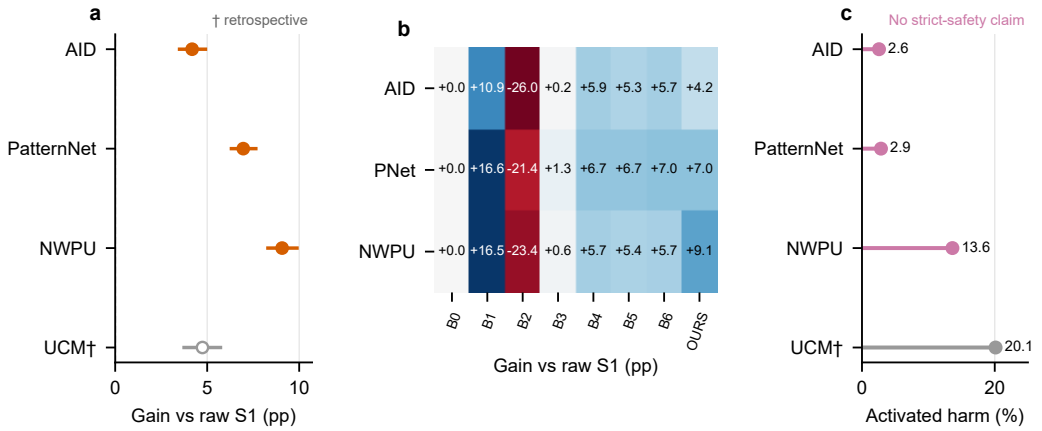


Figure 4: **Historical performance and the adapted-baseline boundary.** (a) V1 gains over raw S1; the open UC Merced marker denotes retrospective evidence. (b) Strong-baseline comparison: adapted S1exceeds V1 on all three historical independent tests. (c) Harmed predictions among activated images.

Table 3: Current V2 ablations on immutable confirmation sets. F1 is Macro-F1; gate is realized activation coverage. Controls use variant-selected development cutoffs. The 1/5–3/5 rows equal 4/5 on both sets and are retained in the full appendix matrix.

Variant	Euro F1	Gate %	C/H	So F1	Gate %	C/H
Adapted S1	0.3168	0.00	0/0	0.0597	0.00	0/0
Unconditional rules	0.2584	85.60	192/263	0.0193	96.06	105/107
Benefit only	0.3711	15.10	78/21	0.0636	6.82	15/4
Single replica 0	0.3754	17.80	88/20	0.0638	16.00	25/1
Replica mean, no vote	0.3731	18.10	83/18	0.0640	5.06	12/0
Dual head, 4/5	0.3721	16.80	78/16	0.0643	8.76	17/0
Dual head, 5/5	0.3724	16.60	78/16	0.0643	8.76	17/0

EuroSAT $n = 1{,}000$; So2Sat $n = 1{,}700$ (stress test). Paired component effects at a shared development target are a different estimand (Figure 5); their exact cutoffs appear in Table 11.

established accuracy benefit. Voting requirements from 1/5 through 4/5 yield identical predictions at the frozen cutoff; 5/5 changes only two EuroSAT gate decisions. The evidence supports selective intervention, but not the necessity of the full component stack for accuracy. Table 3 reports the controls, Figure 5 shows the paired component contrasts, and Table 11 separates variant-selected cutoffs from shared-target contrasts.

4.5 Selective controls, strong baselines, and semantic scope

On EuroSAT, deferral and matched-gate visual routing reach 0.3775/0.3777 Macro-F1, above RuleRouter's 0.3721 (Table 4). So2Sat deferral reaches 0.0861 versus RuleRouter's 0.0643; the matched visual router reaches 0.0635. Dense visual probes reach 0.8358/0.3647 on the two sets, and AID visual correction reaches 0.9709 versus 0.7232. Coverage and correction/harm counts appear in Table 4; these controls bound claims of low-coverage and aggregate-accuracy superiority.

Confusion-driven rules exceed shuffled/random controls on EuroSAT by $2.687/4.255\,pp$ (Holm $p = 0.0033/0.0004$), but not on So2Sat ($+0.124/-0.407\,pp$). The cross-dataset semantic gate fails; these later analyses reuse accessed tests (Appendix Table 21).

Some frozen clauses are anomalous (e.g., "text on image"). So2Sat has dark $[0, 2.8]$ inputs and a prompt without LCZ context. Development-only repair scored 0.14631 OOF Macro-F1, below the frozen 0.147059 gate; no repaired test was opened. Trace cases include corrections and harms, but no human audit (Appendix M).

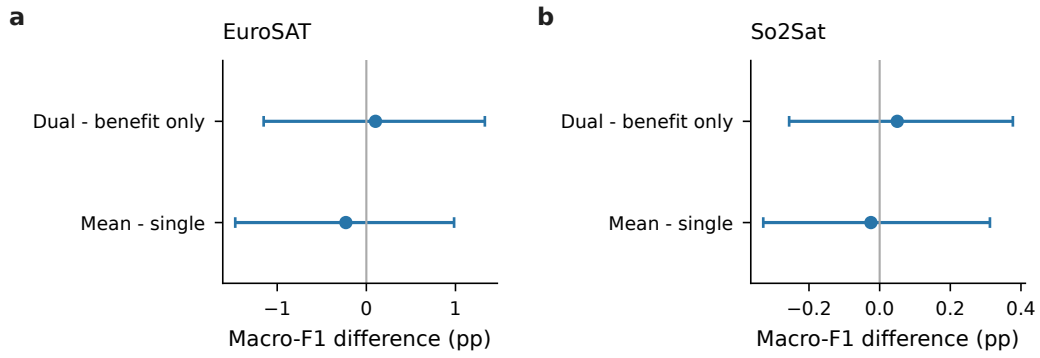


Figure 5: **Paired V2 component effects at a shared development target.** Dual-head 4/5 minus benefit-only, and replica-mean minus canonical single replica, with 95% paired stratified-bootstrap intervals (10,000 draws; $n = 1{,}000/1{,}700$ EuroSAT/So2Sat images). The shared development targets are 15%/10%; realized test coverage differs between methods. These points are not necessarily the variant-selected rows of Table 3. Intervals crossing zero do not establish equivalence.

Table 4: Equal-label selective and dense controls on the same confirmation sets. Each method uses development-only fitting and selection. Coverage is gate activity for routers and changed predictions for dense probes; only the visual matched-gate row has exactly the RuleRouter activation count.

Method	Euro F1	Cov. %	C/H	So F1	Cov. %	C/H
Confidence routing	0.3275	14.20	27/20	0.0593	1.06	0/1
No-text top-2 router	0.3602	15.40	78/14	0.0668	10.65	18/0
Selective defer	0.3775	14.00	90/10	0.0861	12.29	51/5
Visual router (matched gate)	0.3777	16.80	75/18	0.0635	8.76	16/4
RuleRouter	0.3721	16.80	78/16	0.0643	8.76	17/0
S1-logit probe (dense)	0.6122	63.70	362/76	0.2014	88.29	309/67
Visual linear probe (dense)	0.8358	65.40	534/27	0.3647	91.59	571/57

These are additional analyses, not new independent confirmations. The matched visual router uses no rule text. Full paired intervals versus adapted S1 and label/feature budgets are in Appendix G. Point rankings do not establish pairwise significance.

5 Conclusion

RuleRouter applies a confusion-driven rule only when benefit–harm estimates support the proposed correction; otherwise it retains the frozen fallback. Relative to their prespecified comparators, paired Macro-F1 gains are $5.53\,pp$ on EuroSAT, $3.17\,pp$ on AID final-confirm, and $0.46\,pp$ on the low-base-performance So2Sat stress test, with 7.83–16.8% of outputs changed. Selective routing avoids the degradation of unconditional fusion while retaining a textual trace for each intervention.

Dense equal-label visual probes achieve higher aggregate accuracy, while matched low-coverage controls remain competitive. Semantic controls favor confusion-driven rules on EuroSAT but not So2Sat. These results position RuleRouter as a sparse, inspectable intervention mechanism whose value for human auditing remains to be evaluated.

The design separates proposing a correction from accepting it: raw S1 identifies eligible rules, whereas adapted S1 defines the prediction to be replaced. The controls support selective routing without establishing that every router component is necessary. Further evaluation across additional backbones, geographic partitions, and human-audit settings will clarify how well the intervention mechanism and its textual traces transfer.

AI USE STATEMENT

We used generative AI tools to improve the clarity of the manuscript and assist with code implementation. The authors reviewed and revised the AI-assisted text and checked the resulting code for correctness. The authors take responsibility for the final manuscript and accompanying code.

REFERENCES

Anastasios N. Angelopoulos, Stephen Bates, Adam Fisch, Lihua Lei, and Tal Schuster. Conformal risk control. In *International Conference on Learning Representations*, 2024. URL https://openreview.net/forum?id=33XGfHLtZg.

Shuai Bai, Keqin Chen, Xuejing Liu, Jialin Wang, Wenbin Ge, Sibo Song, Kai Dang, Peng Wang, Shijie Wang, Jun Tang, et al. Qwen2.5-vl technical report. *arXiv preprint arXiv:2502.13923*, 2025. URL https://arxiv.org/abs/2502.13923.

Gong Cheng, Junwei Han, and Xiaoqiang Lu. Remote sensing image scene classification: Benchmark and state of the art. *Proceedings of the IEEE*, 105(10):1865–1883, 2017. doi: 10.1109/JPROC.2017.2675998.

Yonatan Geifman and Ran El-Yaniv. Selectivenet: A deep neural network with an integrated reject option. In *Proceedings of the 36th International Conference on Machine Learning*, pp. 2151–2159, 2019. URL https://proceedings.mlr.press/v97/geifman19a.html.

Chuan Guo, Geoff Pleiss, Yu Sun, and Kilian Q. Weinberger. On calibration of modern neural networks. In *Proceedings of the 34th International Conference on Machine Learning*, pp. 1321–1330, 2017. URL https://proceedings.mlr.press/v70/guo17a.html.

Patrick Helber, Benjamin Bischke, Andreas Dengel, and Damian Borth. Eurosat: A novel dataset and deep learning benchmark for land use and land cover classification. *IEEE Journal of Selected Topics in Applied Earth Observations and Remote Sensing*, 12(7):2217–2226, 2019. doi: 10.1109/JSTARS.2019.2918242.

Alon Jacovi and Yoav Goldberg. Towards faithfully interpretable NLP systems: How should we define and evaluate faithfulness? In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pp. 4198–4205. Association for Computational Linguistics, 2020. doi: 10.18653/v1/2020.acl-main.386. URL https://aclanthology.org/2020.acl-main.386/.

Pang Wei Koh, Thao Nguyen, Yew Siang Tang, Stephen Mussmann, Emma Pierson, Been Kim, and Percy Liang. Concept bottleneck models. In *Proceedings of the 37th International Conference on Machine Learning*, pp. 5338–5348, 2020. URL https://proceedings.mlr.press/v119/koh20a.html.

Fan Liu, Delong Chen, Zhangqingyun Guan, Xiaocong Zhou, Jiale Zhu, Qiaolin Ye, Liyong Fu, and Jun Zhou. Remoteclip: A vision language foundation model for remote sensing. *IEEE Transactions on Geoscience and Remote Sensing*, 62:1–16, 2024. doi: 10.1109/TGRS.2024.3390838.

Sachit Menon and Carl Vondrick. Visual classification via description from large language models. In *International Conference on Learning Representations*, 2023. URL https://arxiv.org/abs/2210.07183.

Hussein Mozannar and David Sontag. Consistent estimators for learning to defer to an expert. In *Proceedings of the 37th International Conference on Machine Learning*, pp. 7076–7087, 2020. URL https://proceedings.mlr.press/v119/mozannar20b.html.

Tuomas Oikarinen, Subhro Das, Lam M. Nguyen, and Tsui-Wei Weng. Label-free concept bottleneck models. In *International Conference on Learning Representations*, 2023. URL https://openreview.net/forum?id=FlCg47MNvBA.

Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, et al. Learning transferable visual models from natural language supervision. In *Proceedings of the 38th International Conference on Machine Learning*, pp. 8748–8763, 2021. URL https://proceedings.mlr.press/v139/radford21a.html.

Gui-Song Xia, Jingwen Hu, Fan Hu, Baoguang Shi, Xiang Bai, Yanfei Zhong, Liangpei Zhang, and Xiaoqiang Lu. Aid: A benchmark data set for performance evaluation of aerial scene classification. *IEEE Transactions on Geoscience and Remote Sensing*, 55(7):3965–3981, 2017. doi: 10.1109/TGRS.2017.2685945.

Yi Yang and Shawn Newsam. Bag-of-visual-words and spatial extensions for land-use classification. In *Proceedings of the 18th ACM SIGSPATIAL International Conference on Advances in Geographic Information Systems*, pp. 270–279, 2010. doi: 10.1145/1869790.1869829.

Yue Yang, Artemis Panagopoulou, Shenghao Zhou, Daniel Jin, Chris Callison-Burch, and Mark Yatskar. Language in a bottle: Language model guided concept bottlenecks for interpretable image classification. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 19187–19197, 2023. doi: 10.1109/CVPR52729.2023.01839.

Weixun Zhou, Shawn Newsam, Congmin Li, and Zhenfeng Shao. Patternnet: A benchmark dataset for performance evaluation of remote sensing image retrieval. *ISPRS Journal of Photogrammetry and Remote Sensing*, 145:197–209, 2018. doi: 10.1016/j.isprsjprs.2018.01.004.

Xiao Xiang Zhu, Jingliang Hu, Chunping Qiu, Yilei Shi, Jian Kang, Lichao Mou, Hossein Bagheri, Matthias Haberle, Yuansheng Hua, Rong Huang, Lloyd Hughes, Hao Li, Yao Sun, Guichen Zhang, Shiyao Han, Michael Schmitt, and Yuanyuan Wang. So2sat lcz42: A benchmark data set for the classification of global local climate zones. *IEEE Geoscience and Remote Sensing Magazine*, 8(3): 76–89, 2020. doi: 10.1109/MGRS.2020.2964708.

A COMPACT REPRODUCIBILITY SPECIFICATION

Table 5: **Operating regimes of selective rule intervention and dense equal-label visual adaptation.** The table summarizes the empirical use cases supported by our experiments; it is not an accuracy-superiority claim for RuleRouter.

Criterion	RuleRouter	Dense visual adaptation
Primary objective	Sparse, traceable prediction changes	Maximum aggregate predictive accuracy
Observed comparator changes	7.83–16.8%	Approximately 30–92%
Decision trace	Explicit class-pair textual rule	No matched textual rationale
Non-intervened inputs	Route-specific frozen S1 comparator	Learned visual prediction or correction
Empirical outcome	Positive paired gains with nonzero harm	Higher aggregate Macro-F1
Human auditability	Not yet formally validated	Not designed as a textual rationale

Evidence boundary. A strict-safety route was held after nonzero harm; utility-first V1 then established gains over raw S1; a strong-baseline review found adapted S1 stronger; and V2 was frozen against that comparator before EuroSAT/So2Sat confirm-test access. A later AID final-confirm set was deterministically drawn from a previously unopened reserve and scored once with the already frozen AID model. Failed gates and superseded locks remain archived. UC Merced is retrospective. Figure 6 records this sequence and the corresponding evidence boundary.

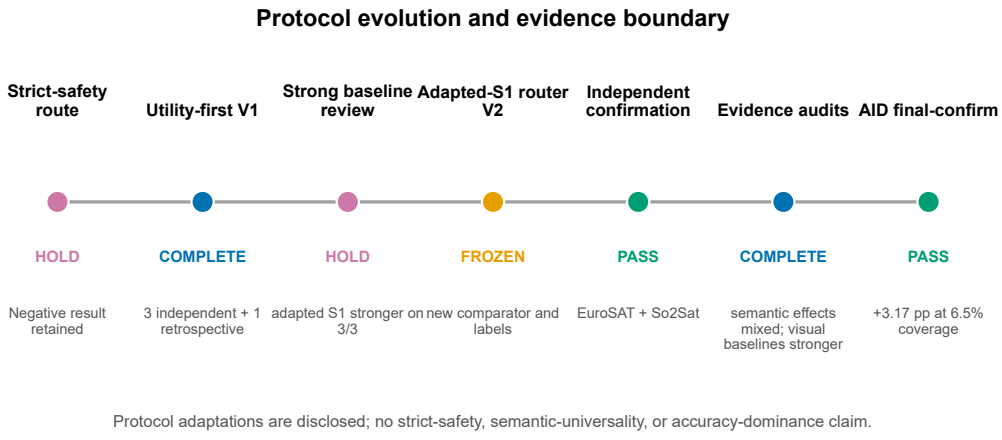


Figure 6: **Protocol evolution and evidence boundary.** Negative safety and strong-baseline findings are retained. The final claim concerns utility-aware improvement on the named confirmatory benchmarks; independent, retrospective, development, and HOLD evidence remain distinct.

Partitions. Native classes use mutually disjoint rule-generation, development, and sealed confirm-test partitions. File identity, hashes, perceptual duplicates, and partition boundaries were audited before scoring. Table 6 covers the original V2 route. The AID final-confirm instead used a separately

Table 6: **Original V2 confirmatory partition sizes.** Per class: rule generation/development/confirm test for EuroSAT and So2Sat.

Dataset	Classes	Rule gen.	Development	Confirm test	Per class
EuroSAT-RGB	10	250	750	1,000	25/75/100
So2Sat-LCZ42	17	425	1,275	1,700	25/75/100

versioned 2,400-image manifest drawn from a never-opened legacy risk partition that was explicitly reassigned before image access.

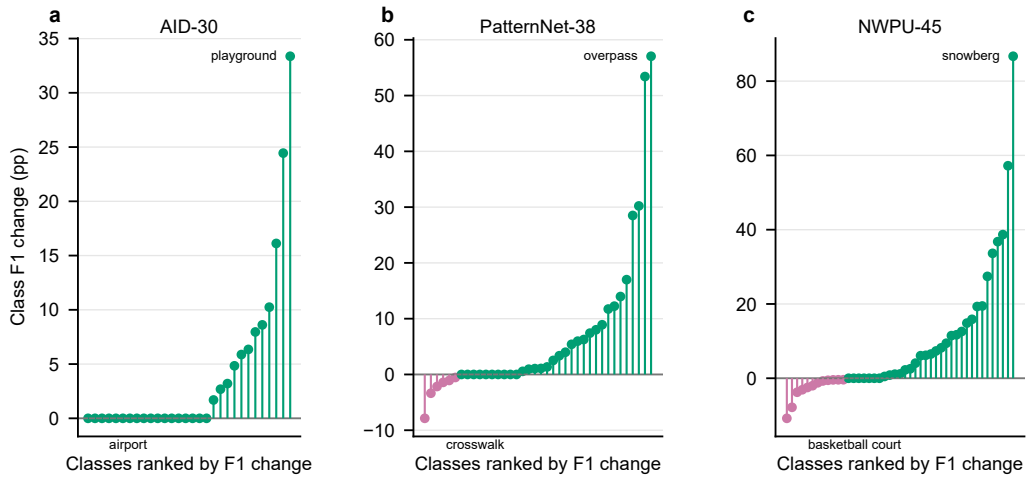


Figure 7: **Class-level heterogeneity of routing effects.** Classes are ranked within each dataset. Green denotes non-negative and magenta denotes negative changes; the most extreme classes are labeled.

Frozen procedure. Nested out-of-fold splits fit T and zero-sum class bias on development images. Separately, rule-generation images define one directed confusion edge per source class and three representatives per confusion group. Rules contain at most two clauses, class names are forbidden inside clauses, and at most four rules are retained per edge. Five bootstrap L1-logistic benefit/harm routers operate on the $3C + 4$ features in Section 3. The V2 activation rule requires a changed proposal, 4-of-5 agreement, and $u > 0$ with $u \geq \tau$, where $u = p_{B,(2)} - p_{H,(4)}$. Appendix E specifies the distinct AID comparator accounting.

Prompts and fixed settings. VLM inference ran on one NVIDIA A100-PCIE-40GB GPU. For EuroSAT and So2Sat, the System-1 template is `Is this satellite image best described as {class_name}? Answer only Yes or No.`; the separately frozen AID route uses `This is a satellite land-use image. Is its correct class '{class_name}'? Answer only Yes or No.` All routes score clauses with `Does this satellite image visibly show the following property? {clause} Answer only Yes or No.` The frozen model uses deterministic decoding, bfloat16, and batch size 16. Seeds are 42, 123, 2025, 3407, and 10086. The EuroSAT/So2Sat coverage grid is $\{0, 1, 2.5, 5, 7.5, 10, 15\}\%$, with cutoffs 0.310770783333017 and 0.5228248143199375; AID preserves its separately frozen 156-image activation budget. Confidence intervals use 10,000 paired stratified-bootstrap replicates, with seed 20270905 for EuroSAT/So2Sat and seed 20270909 for AID final-confirm.

B Class-level heterogeneity

Figure 7 reports per-class F1 changes for the historical independent tests. Dataset-level gains coexist with negative class-level effects, reinforcing the need to report native-class coverage and not hide heterogeneity behind a single aggregate.

C Historical V1 results and the adapted-baseline reversal

C.1 Supervised logit adaptation reverses the historical conclusion

Relative to raw S1, the historical V1 router appears consistently effective: gains are $+4.180\,pp$ on AID, $+6.961\,pp$ on PatternNet, and $+9.071\,pp$ on NWPU (Table 2). The equal-dataset mean is $+6.74\,pp$, and all three Holm-adjusted exact McNemar tests are significant. UC Merced adds a retrospective $+4.742\,pp$ result but is excluded from the independent average.

Supervised temperature-and-class-bias adaptation changes that conclusion. Development-fitted adapted S1 exceeds V1 by 6.766, 9.614, and 7.426 points on AID, PatternNet, and NWPU, respectively (Figure 4b). The historical evidence therefore supports V1 over raw S1, confidence routing, and unconditional rules, but not over the strongest adapted baseline. This reversal motivated the stronger comparator, benefit–harm targets, and the subsequent frozen confirmation protocol.

C.2 Selectivity prevents destructive rule application

The largest historical ablation effect comes from routing itself (Figure 9). Selective routing exceeds unconditional rule fusion by 30.22, 28.37, and 32.49 Macro-F1 points on the historical AID, PatternNet, and NWPU tests. The isolated effects of cross-replica sharing, dual heads, and 4-of-5 consensus are dataset-dependent. For example, dual-head consensus relative to benefit-only changes Macro-F1 by -1.76, $+0.28$, and $+3.34\,pp$, while consensus relative to a shared router changes it by -1.55, -0.03, and $+3.37\,pp$. These V1 comparisons are not substitutes for the current V2 component evidence in Section 4.4.

The risk–coverage sweeps in Figure 8 explain this heterogeneity. Increasing activation exposes more potential corrections but can also raise the fraction of harmed activated samples. The development-selected point is therefore part of the method, not a presentation-only threshold.

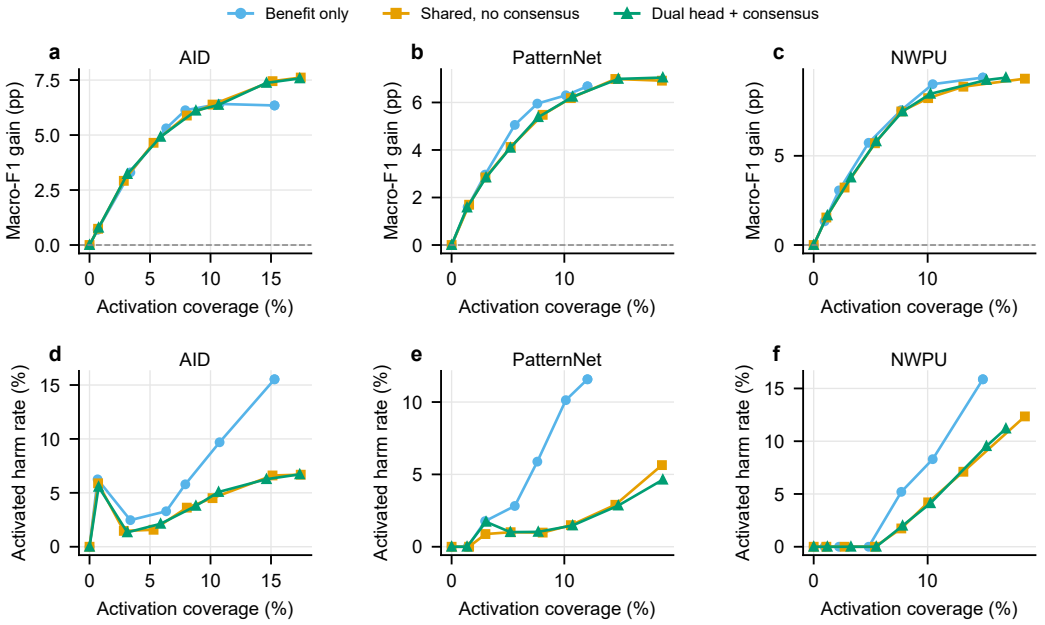


Figure 8: **Historical V1 utility–harm trade-offs.** Top: Macro-F1 gain over raw S1. Bottom: harmed predictions among activated samples. Curves compare benefit-only, shared routing without consensus, and dual-head 4-of-5 consensus. The current V2 curves against adapted S1 are in Appendix F.

D Reading the current V2 comparisons

Main-text Table 3 reports variant-selected development cutoffs, whereas Figure 5 compares components at shared development targets of 15% on EuroSAT and 10% on So2Sat. Realized test coverage need not match. For example, So2Sat benefit-only activates 6.82% under its selected cutoff but 9.00% under the shared target. Table 11 maps the exact policies; the full grid and paired intervals follow in Appendix F.

The matched semantic controls use 150 EuroSAT and 170 So2Sat activations, distinct from the primary policies' 168 and 149. Their complete results and multiplicity adjustment are reported with the semantic arm tables. Image-linked traces and deterministic case selection are documented once, in Appendix M. These cases illustrate the available record, not a human evaluation.

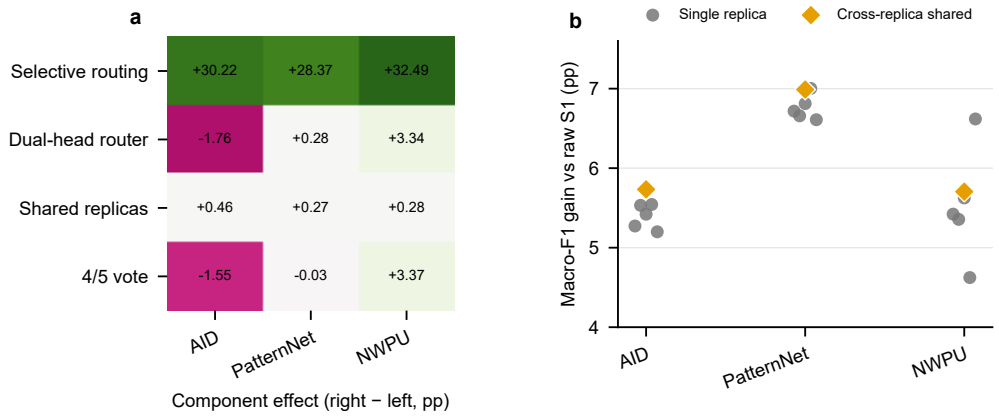


Figure 9: **Historical V1 component effects and cross-replica behavior.** Selective routing prevents the large degradation caused by unconditional rules. Sharing, dual heads, and consensus have dataset-dependent effects. Current V2 paired component contrasts are reported separately in Appendix F.

E Confirmation metrics and comparator accounting

E.1 What counts as an intervention?

We distinguish a router's gate indicator a_i from a changed output $d_i = \mathbb{1}[\hat{y}_i \neq \hat{y}_{\mathrm{ref},i}]$. Gate coverage is $n^{-1}\sum_i a_i$, whereas change coverage is $n^{-1}\sum_i d_i$. Corrections and harms are defined relative to a named reference prediction, not relative to whether a gate fired:

$$C = \sum_i \mathbb{1}[\hat{y}_{\mathrm{ref},i} \neq y_i,\ \hat{y}_i = y_i], \tag{12}$$

$$H = \sum_i \mathbb{1}[\hat{y}_{\mathrm{ref},i} = y_i,\ \hat{y}_i \neq y_i]. \tag{13}$$

Changed outputs that are neither corrections nor harms are wrong-to-wrong changes. They are not prediction-unchanged cases. The accuracy difference satisfies $(C - H)/n$; Macro-F1 need not follow this identity because it weights classes rather than individual images.

Table 7: Detailed independent-confirmation metrics. Accuracy and coverage are percentages; Macro-F1 is on a 0–1 scale.

Dataset	S1 acc.	Final acc.	Gate n	Changed n	C	H
EuroSAT	32.90	39.10	168	168	78	16
AID final-confirm	70.96	74.17	156	188	104	27
So2Sat	6.71	7.71	149	149	17	0

E.2 AID reporting comparator versus internal fallback

The AID final-confirm archive contains both the reporting S1 prediction and the internal S1 prediction used by the frozen router. These are not identical. Table 8 reconstructs the distinction from the same 2,400 per-image records, without changing predictions, models, or the locked primary endpoint. Against reporting S1, 151 changed outputs occur on the 156 gate-active images and 37 occur on gate-inactive images. Thus the full pipeline changes 188 outputs (7.83%), while the gate activates 156 times (6.5%). Across all images, the 104 corrections and 27 harms reproduce the primary accuracy gain of $77/2400 = 3.2083$ percentage points.

Against the internal fallback, gate-inactive predictions are unchanged and the 156 active interventions produce 97 corrections and six harms. Against reporting S1, the gate-active subset instead has 93 corrections and ten harms, and the inactive subset has 11 corrections and 17 harms. Consequently, $27/156$ mixes an all-image harm numerator with a gate-only denominator and is not reported as activated-sample harm. The corresponding conditional rates are $10/156$ against reporting S1 and $6/156$ against the internal fallback; these describe different references. The primary $+3.172$-point

Table 8: AID comparator reconciliation from the same 2,400 frozen per-image outputs. Counts are descriptive and do not change the locked predictions or primary comparison.

Reference	Gate	Images	Changed	Corrected	Harmed
reporting S1	active	156	151	93	10
reporting S1	inactive	2244	37	11	17
internal fallback	active	156	156	97	6
internal fallback	inactive	2244	0	0	0

Macro-F1 comparison and its interval remain those against reporting S1. They must not be attributed entirely to the isolated rule gate.

Scope of this clarification. This is a manuscript-side accounting reconciliation, not a new experiment or a revised selection rule. Original result locks are preserved. It also explains why a generic "unchanged outside the gate" statement applies to the internal fallback but not to every externally reported comparator.

F Complete V2 component and voting results

The following analyses reuse the immutable EuroSAT and So2Sat confirmation outputs. They are additional, retrospective component analyses, not a second independent test. No variant replaces the frozen primary model. Development OOF outputs determine cutoffs; inspecting these curves does not authorize a new test-selected operating point.

Table 9: V2 ablations at variant-selected development cutoffs on EuroSAT. Benefit-only, single-replica, and mean-router rows use their separately selected OOF operating points, not necessarily the common target used for paired component contrasts.

Variant	Macro-F1	Gain [95% CI]	Gate %	C/H
Adapted S1	0.3168	+0.00 $[+0.00, +0.00]$	0.00	0/0
Unconditional rules	0.2584	-5.85 $[-9.58, -2.32]$	85.60	192/263
Benefit only	0.3711	+5.42 $[+3.64, +7.16]$	15.10	78/21
Single replica 0	0.3754	+5.86 $[+3.99, +7.65]$	17.80	88/20
Replica mean, no vote	0.3731	+5.63 $[+3.91, +7.24]$	18.10	83/18
Dual head, 1/5	0.3721	+5.53 $[+3.90, +7.06]$	16.80	78/16
Dual head, 2/5	0.3721	+5.53 $[+3.90, +7.06]$	16.80	78/16
Dual head, 3/5	0.3721	+5.53 $[+3.90, +7.06]$	16.80	78/16
Dual head, 4/5	0.3721	+5.53 $[+3.90, +7.06]$	16.80	78/16
Dual head, 5/5	0.3724	+5.55 $[+3.94, +7.09]$	16.60	78/16
Full frozen model	0.3721	+5.53 $[+3.90, +7.06]$	16.80	78/16

A8 (4/5) and A10 (full) are the same frozen policy. Table 11 lists the cutoffs and target budgets that distinguish this table from the component contrasts. The five models are bootstrap replicas, not independent test replications.

The single-replica control uses the prespecified canonical replica, not the best test-performing replica. "Shared mean, no vote" averages head outputs without a consensus count. Voting controls require one through five agreeing replicas under the archived selection procedure. The full model and the 4/5 row are identical predictions; their repetition is a consistency check, not replication of the experiment.

Distinct operating points, not conflicting measurements. Tables 9 and 10 use variant-selected development cutoffs. The first two paired component contrasts below instead use a common development target: 15% on EuroSAT and 10% on So2Sat. On So2Sat, benefit-only changes from target 7.5% (cutoff 0.647764, F1 0.063605, gate 6.82%) to target 10% (cutoff 0.618607, F1 0.063800, gate 9.00%). Mean routing changes from target 5% (cutoff 0.642529, F1 0.064031, gate 5.06%) to target 10% (cutoff 0.570426, F1 0.063586, gate 8.24%). Hence the component deltas must not be computed by subtracting the differently selected ablation rows. All original numerical results are retained; this revision repairs their labels and estimand definitions.

Table 10: V2 ablations at variant-selected development cutoffs on So2Sat. Benefit-only, single-replica, and mean-router rows use their separately selected OOF operating points, not necessarily the common target used for paired component contrasts.

Variant	Macro-F1	Gain [95% CI]	Gate %	C/H
Adapted S1	0.0597	+0.00 $[+0.00, +0.00]$	0.00	0/0
Unconditional rules	0.0193	-4.04 $[-5.20, -2.90]$	96.06	105/107
Benefit only	0.0636	+0.39 $[+0.02, +0.75]$	6.82	15/4
Single replica 0	0.0638	+0.41 $[+0.02, +0.79]$	16.00	25/1
Replica mean, no vote	0.0640	+0.43 $[+0.15, +0.74]$	5.06	12/0
Dual head, 1/5	0.0643	+0.46 $[+0.17, +0.77]$	8.76	17/0
Dual head, 2/5	0.0643	+0.46 $[+0.17, +0.77]$	8.76	17/0
Dual head, 3/5	0.0643	+0.46 $[+0.17, +0.77]$	8.76	17/0
Dual head, 4/5	0.0643	+0.46 $[+0.17, +0.77]$	8.76	17/0
Dual head, 5/5	0.0643	+0.46 $[+0.17, +0.77]$	8.76	17/0
Full frozen model	0.0643	+0.46 $[+0.17, +0.77]$	8.76	17/0

A8 (4/5) and A10 (full) are the same frozen policy. Table 11 lists the cutoffs and target budgets that distinguish this table from the component contrasts. The five models are bootstrap replicas, not independent test replications.

Table 11: Operating-point reconciliation. Selected denotes the per-variant OOF choice in the ablation matrix; shared target denotes the first two paired component contrasts. All values are recovered from the original locked risk–coverage grid, without retuning or altering predictions.

Method	Selection	Dev. target %	Cutoff	Test F1	Test gate %
EuroSAT					
Benefit only	Selected	15.00	0.697804	0.371087	15.10
Benefit only	Shared target	15.00	0.697804	0.371087	15.10
Single replica	Selected	15.00	0.445231	0.375433	17.80
Single replica	Shared target	15.00	0.445231	0.375433	17.80
Mean, no vote	Selected	15.00	0.418208	0.373136	18.10
Mean, no vote	Shared target	15.00	0.418208	0.373136	18.10
So2Sat					
Benefit only	Selected	7.50	0.647764	0.063605	6.82
Benefit only	Shared target	10.00	0.618607	0.063800	9.00
Single replica	Selected	10.00	0.610077	0.063833	16.00
Single replica	Shared target	10.00	0.610077	0.063833	16.00
Mean, no vote	Selected	5.00	0.642529	0.064031	5.06
Mean, no vote	Shared target	10.00	0.570426	0.063586	8.24

F.1 Paired effects of individual components

Figure 5 visualizes the first two paired contrasts. Shared development targets do not imply equal realized test activation; Tables 12 and 13 report the exact rates.

The large contrast is between selective and unconditional application. Dual-head versus benefit-only differences are much smaller, and the paired intervals cross zero. This supports a benefit/harm accounting design but not an independent accuracy-improvement claim for the harm head. Likewise, 4/5 is a prespecified operating point, not a universally optimal vote count. Overlapping intervals are not used as a formal equality test; the paired contrasts themselves are the relevant estimands.

Table 12: Paired component contrasts on EuroSAT. The first two rows use a common 15% development target, not the variant-selected operating points in Table 9. Realized confirmation coverage is not matched.

Contrast	Delta (pp)	95% CI (pp)	Left/right gate %
4/5 minus benefit only (shared target)	+0.10	$[-1.15, +1.33]$	15.10 / 16.80
Mean minus single (shared target)	-0.23	$[-1.47, +0.98]$	17.80 / 18.10
full model minus unconditional rule fusion	+11.38	$[+8.20, +14.66]$	85.60 / 16.80
4 minus 1	+0.00	$[+0.00, +0.00]$	16.80 / 16.80
4 minus 2	+0.00	$[+0.00, +0.00]$	16.80 / 16.80
4 minus 3	+0.00	$[+0.00, +0.00]$	16.80 / 16.80
4 minus 5	-0.03	$[-0.08, +0.00]$	16.60 / 16.80

Differences are right minus left. The remaining rows compare the frozen full policy with unconditional fusion or vary only the required vote count at its fixed cutoff. Table 11 resolves the distinct estimands.

Table 13: Paired component contrasts on So2Sat. The first two rows use a common 10% development target, not the variant-selected operating points in Table 10. Realized confirmation coverage is not matched.

Contrast	Delta (pp)	95% CI (pp)	Left/right gate %
4/5 minus benefit only (shared target)	+0.05	$[-0.26, +0.38]$	9.00 / 8.76
Mean minus single (shared target)	-0.02	$[-0.33, +0.31]$	16.00 / 8.24
full model minus unconditional rule fusion	+4.50	$[+3.37, +5.65]$	96.06 / 8.76
4 minus 1	+0.00	$[+0.00, +0.00]$	8.76 / 8.76
4 minus 2	+0.00	$[+0.00, +0.00]$	8.76 / 8.76
4 minus 3	+0.00	$[+0.00, +0.00]$	8.76 / 8.76
4 minus 5	+0.00	$[+0.00, +0.00]$	8.76 / 8.76

Differences are right minus left. The remaining rows compare the frozen full policy with unconditional fusion or vary only the required vote count at its fixed cutoff. Table 11 resolves the distinct estimands.

F.2 Coverage and harm across frozen development choices

A target development coverage does not guarantee the same confirmation coverage under distribution shift. We therefore retain the realized rates instead of relabeling a test point with its development target. At zero activation, conditional harm has no denominator; it is undefined rather than evidence of a perfectly safe nontrivial policy. Macro-F1 and harm must be read jointly because additional corrections can coexist with additional damage to previously correct predictions.

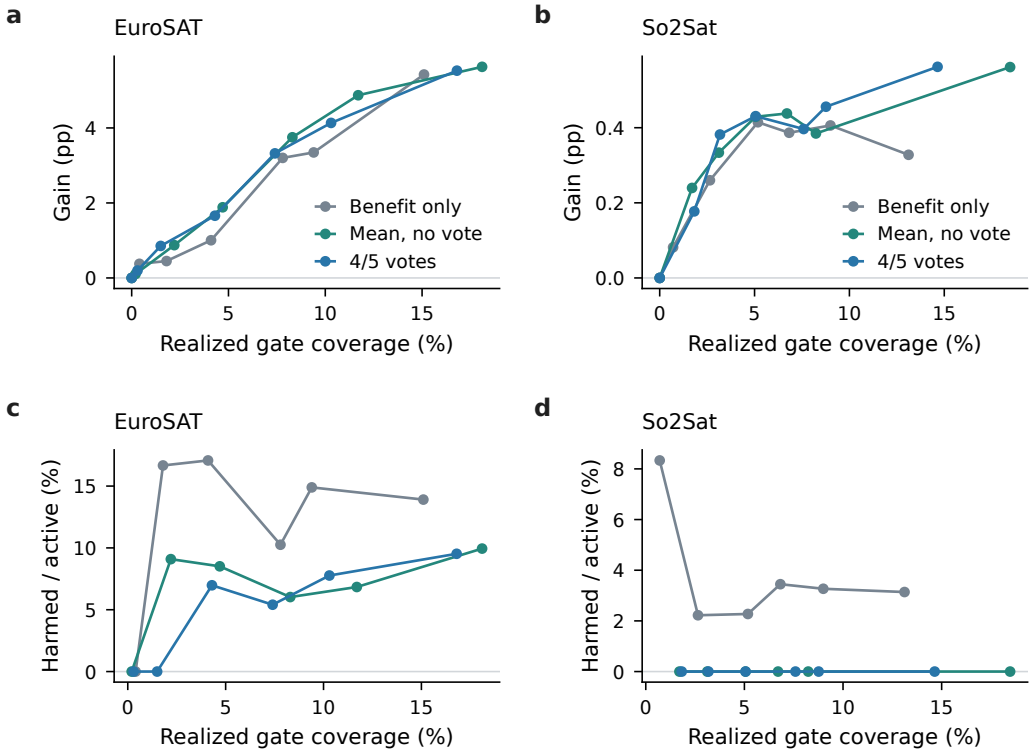


Figure 10: **V2 utility and harm across development-derived cutoffs.** Each point uses a cutoff obtained from the prespecified development coverage grid; the horizontal coordinate is realized confirmation coverage. Top: Macro-F1 gain over adapted S1. Bottom: harms among gate-active images. These are descriptive sensitivity curves, not instructions to retune on confirmation outcomes. The exact point values are supplied in `coverage.csv`.

G Equal-label supervised comparisons

G.1 Features, label budgets, and selection

The EuroSAT/So2Sat controls use the same 750/1,275 development labels and the five seeds in Section 4.1. Trainable controls use five-fold development evaluation and the frozen seven-value regularization grid $\{0.01, 0.03, 0.1, 0.3, 1, 3, 10\}$. The grid is a within-protocol fairness constraint, not a claim that every historical AID protocol used seven trials. Test labels do not fit, rank, or select models.

S1-logit controls use cached class evidence without visual embeddings or rule text. Visual controls use the final frozen Qwen visual-tower output after spatial merge, averaged over image tokens and L2-normalized. The embeddings are stored in float32; the tower runs in its frozen bfloat16 configuration and receives no text prompt. The visual linear probe uses only this embedding. Visual correction also receives adapted-S1 features. Visual proposal routers have benefit/harm heads but no clause score or text. Their matched-coverage version ranks by the same frozen scores and retains exactly the full model's activation count without consulting test labels.

Table 14: EuroSAT equal-label cached S1-feature comparisons. This is additional analysis of a previously accessed, immutable confirmation set; fitting and selection use development labels only.

Method	Macro-F1	Gain [95% CI]	Change/gate %	C/H
Adapted S1	0.3168	+0.00 [+0.00, +0.00]	0.00	0/0
Temperature only	0.2872	-2.96 [−5.50, −0.48]	44.70	107/90
S1-logit probe	0.6122	+29.54 [+26.04, +32.95]	63.70	362/76
S1 correction classifier	0.5635	+24.67 [+21.09, +28.04]	67.10	333/94
No-text top-2 correction router	0.3602	+4.34 [+2.77, +5.93]	15.40	78/14
Confidence routing	0.3275	+1.06 [−0.46, +2.58]	14.20	27/20
Margin routing	0.3152	-0.16 [−1.00, +0.69]	4.70	5/8
Selective defer	0.3775	+6.07 [+4.47, +7.73]	14.00	90/10
RuleRouter	0.3721	+5.53 [+3.92, +7.13]	16.80	78/16

For dense classifiers, coverage counts changed predictions relative to adapted S1; for routers, it counts gate activity. Visual-control CIs use seed 20270908, so the repeated full-model interval may differ slightly from the original primary interval (seed 20270905).

Table 15: So2Sat equal-label cached S1-feature comparisons. This is additional analysis of a previously accessed, immutable confirmation set; fitting and selection use development labels only.

Method	Macro-F1	Gain [95% CI]	Change/gate %	C/H
Adapted S1	0.0597	+0.00 [+0.00, +0.00]	0.00	0/0
Temperature only	0.0108	-4.90 [−6.01, −3.79]	72.82	71/94
S1-logit probe	0.2014	+14.17 [+12.22, +16.10]	88.29	309/67
S1 correction classifier	0.1739	+11.42 [+9.61, +13.17]	86.18	268/68
No-text top-2 correction router	0.0668	+0.71 [+0.21, +1.28]	10.65	18/0
Confidence routing	0.0593	-0.04 [−0.14, +0.02]	1.06	0/1
Margin routing	0.0562	-0.35 [−0.92, +0.18]	8.47	10/12
Selective defer	0.0861	+2.64 [+1.70, +3.60]	12.29	51/5
RuleRouter	0.0643	+0.46 [+0.16, +0.77]	8.76	17/0

For dense classifiers, coverage counts changed predictions relative to adapted S1; for routers, it counts gate activity. Visual-control CIs use seed 20270908, so the repeated full-model interval may differ slightly from the original primary interval (seed 20270905).

G.2 Visual-feature baselines

The visual linear probe and visual correction classifier are substantially more accurate overall. At matched coverage, the visual router is slightly better on EuroSAT and slightly worse on So2Sat. Neither observation supports a general superiority claim for natural-language rules under a low-coverage constraint. A textual trace is an architectural output; its usefulness to a human evaluator was not measured by these experiments.

Table 16: EuroSAT equal-label visual-feature comparisons. This is additional analysis of a previously accessed, immutable confirmation set; fitting and selection use development labels only.

Method	Macro-F1	Gain [95% CI]	Change/gate %	C/H
Adapted S1	0.3168	+0.00 [+0.00, +0.00]	0.00	0/0
Visual linear probe	0.8358	+51.89 [+48.76, +54.99]	65.40	534/27
Visual correction classifier	0.8326	+51.58 [+48.40, +54.72]	66.00	533/29
Visual proposal router	0.3748	+5.80 [+4.11, +7.54]	14.60	66/12
Visual router, matched	0.3777	+6.09 [+4.26, +7.99]	16.80	75/18
RuleRouter	0.3721	+5.53 [+3.93, +7.13]	16.80	78/16

For dense classifiers, coverage counts changed predictions relative to adapted S1; for routers, it counts gate activity. Visual-control CIs use seed 20270908, so the repeated full-model interval may differ slightly from the original primary interval (seed 20270905).

Table 17: So2Sat equal-label visual-feature comparisons. This is additional analysis of a previously accessed, immutable confirmation set; fitting and selection use development labels only.

Method	Macro-F1	Gain [95% CI]	Change/gate %	C/H
Adapted S1	0.0597	+0.00 [+0.00, +0.00]	0.00	0/0
Visual linear probe	0.3647	+30.49 [+28.10, +32.81]	91.59	571/57
Visual correction classifier	0.3595	+29.97 [+27.60, +32.25]	91.94	564/60
Visual proposal router	0.0632	+0.35 [+0.00, +0.70]	5.94	12/2
Visual router, matched	0.0635	+0.37 [−0.02, +0.76]	8.76	16/4
RuleRouter	0.0643	+0.46 [+0.16, +0.77]	8.76	17/0

For dense classifiers, coverage counts changed predictions relative to adapted S1; for routers, it counts gate activity. Visual-control CIs use seed 20270908, so the repeated full-model interval may differ slightly from the original primary interval (seed 20270905).

G.3 AID FINAL-CONFIRM BASELINES

Table 18: All seven frozen AID final-confirm methods, evaluated with the same development-label budget. Delta and C/H are relative to reporting S1. Gate/change coverage follows each method definition.

Method	Macro-F1	Gain [95% CI]	Gate/change %	C/H
Reporting S1	0.6915	+0.00 [+0.00, +0.00]	0.00	0/0
S1-logit probe	0.9373	+24.58 [+22.92, +26.35]	29.38	605/52
Visual linear probe	0.9658	+27.43 [+25.75, +29.22]	29.96	656/38
Visual correction classifier	0.9709	+27.94 [+26.29, +29.72]	29.71	661/30
Confidence routing	0.6833	-0.82 [−1.38, −0.29]	2.50	8/27
Selective defer	0.8278	+13.63 [+12.36, +15.00]	15.00	327/12
RuleRouter	0.7232	+3.17 [+2.26, +4.15]	6.50	104/27

RuleRouter uses 156 gate activations; its total reporting-comparator changes are 188. Dividing all 27 harms by 156 would mix the comparator and gate denominators, so that ratio is not interpreted as an activated-sample harm rate here.

All AID baselines use the same authorized development labels. Their separate frozen regularization grid is $\{0.01, 0.1, 1, 10, 100\}$ with five-fold development selection. Final-confirm labels are not a training resource. The visual baseline advantage is disclosed rather than excluded on the grounds that its changes are denser. A fair interpretation is that the preferred method depends on whether aggregate accuracy or constrained, traceable intervention is the objective; these experiments do not establish an optimal trade-off for every application.

G.4 PROBABILITY QUALITY

Table 19: Probability metrics for methods that emit normalized multiclass probabilities. ECE uses 15 bins. Hard routing policies have no assigned ECE/NLL/Brier in this comparison.

Dataset / method	ECE	NLL	Brier
EuroSAT / Adapted S1	0.0473	1.8201	0.7679
EuroSAT / Visual linear probe	0.0416	0.4936	0.2401
EuroSAT / Visual correction classifier	0.0466	0.4858	0.2377
So2Sat / Adapted S1	0.0060	2.8355	0.9413
So2Sat / Visual linear probe	0.1169	2.0531	0.7892
So2Sat / Visual correction classifier	0.1276	2.0584	0.7893

Temperature-only scaling preserves argmax predictions. Temperature plus class bias can change them and is supervised adaptation, even when the same optimization also improves probability quality. NLL, Brier score, and ECE evaluate probability distributions, whereas the main endpoint evaluates hard labels. We do not invent calibrated probability vectors for hard routing policies solely to populate this table.

H MATCHED SEMANTIC CONTROLS

The four text arms are confusion-driven rules, ordinary category descriptions, source–target-shuffled rules, and random-matched rules. Their rule/clause structure and scoring budgets are matched under the archived protocol. These controls test an attribution claim; they do not replace the original confirmation operating point. They use exactly 150 EuroSAT activations and 170 So2Sat activations, rather than the primary 168 and 149. The no-text supervised alternatives are reported separately in Appendix G.

Table 20: All matched text arms. EuroSAT activates exactly 150/1,000 images and So2Sat 170/1,700. These are controlled attribution runs, distinct from the primary frozen operating points.

Text arm	Macro-F1	Delta (pp)	Gate %	C/H	H/active %
EuroSAT					
Confusion-driven	0.3737	+5.69	15.00	76/15	10.00
Category descriptions	0.3430	+2.62	15.00	49/20	13.33
Source–target shuffled	0.3469	+3.00	15.00	61/22	14.67
Random-matched	0.3312	+1.44	15.00	45/29	19.33
So2Sat					
Confusion-driven	0.0644	+0.47	10.00	15/0	0.00
Category descriptions	0.0715	+1.17	10.00	24/1	0.59
Source–target shuffled	0.0632	+0.34	10.00	12/0	0.00
Random-matched	0.0685	+0.88	10.00	18/0	0.00

Table 21: The four prespecified semantic contrasts. Intervals are paired stratified-bootstrap intervals; one-sided empirical bootstrap tail probabilities use a plus-one correction and Holm adjustment across these four tests.

Dataset / comparator	Delta (pp)	95% CI (pp)	Holm p
EuroSAT / shuffled	+2.69	$[+1.00, +4.39]$	0.0033
EuroSAT / random	+4.25	$[+2.52, +5.99]$	0.0004
So2Sat / shuffled	+0.12	$[-0.34, +0.59]$	0.5963
So2Sat / random	-0.41	$[-0.96, +0.14]$	0.9292

Tables 20 and 21 retain all arms, paired contrasts, and multiplicity-adjusted tests.

The four primary comparisons form one family: two controls on each of two datasets. The archived one-sided empirical bootstrap tail probability uses a plus-one correction, followed by Holm adjustment at familywise $\alpha = 0.05$. This is an empirical bootstrap procedure, not an exact permutation test. The gate requires positive paired differences and adjusted significance for all four comparisons. Because it fails, our cross-dataset claim is limited to *VLM-derived rule features*.

What semantic controls do not establish. They do not measure whether an individual clause is factually true, whether it faithfully explains the VLM's internal reasoning, or whether a person can use it to identify an incorrect intervention. The rule is an explicit computational trace, not a validated human explanation. There is no blinded human annotation result in this paper. Generated clauses should therefore remain inspectable, including semantically weak ones, rather than being cleaned up after the outcomes are known.

I NATIVE-CLASS CONFIRMATION RESULTS

I.1 EUROSAT

Table 22: All native-class F1 scores on EuroSAT confirmation images. Every class has 100 images; delta is RuleRouter minus adapted S1. The visual linear probe is an equal-label comparison.

Native class	Adapted S1	RuleRouter	Delta (pp)	Visual probe
AnnualCrop	0.0769	0.0909	+1.40	0.8317
Forest	0.3388	0.3564	+1.76	0.9216
HerbaceousVegetation	0.1106	0.3750	+26.44	0.7940
Highway	0.5171	0.5248	+0.77	0.6927
Industrial	0.5437	0.5463	+0.27	0.8846
Pasture	0.0594	0.2727	+21.33	0.8421
PermanentCrop	0.1761	0.1333	-4.28	0.7475
Residential	0.5565	0.5575	+0.10	0.9200
River	0.4211	0.4255	+0.45	0.7835
SeaLake	0.3682	0.4386	+7.04	0.9400

I.2 SO2SAT

Table 23: All native-class F1 scores on So2Sat confirmation images. Every class has 100 images; delta is RuleRouter minus adapted S1. The visual linear probe is an equal-label comparison.

Native class	Adapted S1	RuleRouter	Delta (pp)	Visual probe
Compact high-rise	0.0164	0.0164	+0.00	0.4257
Compact mid-rise	0.0000	0.0000	+0.00	0.3316
Compact low-rise	0.0342	0.0342	+0.00	0.1865
Open high-rise	0.0829	0.0829	+0.00	0.3187
Open mid-rise	0.0847	0.0847	+0.00	0.1954
Open low-rise	0.0000	0.0000	+0.00	0.2328
Lightweight low-rise	0.0000	0.0000	+0.00	0.3777
Large low-rise	0.0182	0.0182	+0.00	0.3959
Sparsely built	0.0898	0.1007	+1.09	0.3237
Heavy industry	0.1382	0.1382	+0.00	0.3302
Dense trees	0.0000	0.0000	+0.00	0.5000
Scattered trees	0.0568	0.1234	+6.66	0.2247
Bush or scrub	0.0000	0.0000	+0.00	0.3610
Low plants	0.0174	0.0174	+0.00	0.4259
Bare rock or paved	0.0952	0.0952	+0.00	0.3529
Bare soil or sand	0.2995	0.2995	+0.00	0.5182
Water	0.0824	0.0824	+0.00	0.6986

I.3 AID FINAL-CONFIRM

Table 24: All 30 native-class results on AID final-confirm. Support is not equal across classes because the previously sealed partition was preserved. Delta is relative to reporting S1; the visual column is the correction classifier.

Class	Support	S1	RuleRouter	Delta (pp)	Visual
airport	87	0.9942	0.9942	+0.00	0.9942
bare land	74	0.7531	0.7578	+0.47	0.9863
baseball field	50	0.8980	0.9091	+1.11	0.9600
beach	96	0.9741	0.9741	+0.00	1.0000
bridge	87	0.8790	0.8790	+0.00	0.9942
center	62	0.0370	0.1688	+13.18	0.9531
church	58	0.6190	0.6118	-0.73	0.9412
commercial	84	0.5464	0.5905	+4.40	0.9213

Continued on next page

Table 24 continued

Class	Support	S1	RuleRouter	Delta (pp)	Visual
dense residential	99	0.8372	0.8372	+0.00	0.9744
desert	72	0.7634	0.7727	+0.94	0.9931
farmland	89	0.7479	0.7386	-0.93	0.9944
forest	60	0.7883	0.7714	-1.69	0.9833
industrial	94	0.6069	0.6537	+4.68	0.9783
meadow	67	0.5981	0.5926	-0.55	0.9701
medium residential	69	0.7218	0.7143	-0.75	0.9571
mountain	82	0.9427	0.9427	+0.00	1.0000
park	84	0.6825	0.7660	+8.35	0.9341
parking	94	0.9347	0.9394	+0.47	0.9947
playground	89	0.3670	0.7215	+35.45	0.9727
pond	101	0.8959	0.8991	+0.32	0.9754
port	92	0.9508	0.9508	+0.00	0.9891
railway station	63	0.8037	0.8000	-0.37	0.9606
resort	69	0.7321	0.7544	+2.22	0.9265
river	98	0.8098	0.8039	-0.58	0.9796
school	72	0.3878	0.3542	-3.36	0.9155
sparse residential	72	0.6423	0.6016	-4.07	1.0000
square	79	0.4000	0.6438	+24.38	0.9281
stadium	69	0.5872	0.7977	+21.05	0.9706
storage tanks	87	0.5000	0.4394	-6.06	0.9884
viaduct	101	0.3443	0.3167	-2.76	0.9901

Macro-F1 gives equal weight to classes even when support differs. We do not remove difficult classes, remap ontologies to improve scores, or average retrospective UC Merced results into independent confirmation. The class-level tables expose heterogeneity that an aggregate gain hides.

J So2Sat input diagnosis and failed development repair

The original So2Sat rendering uses the correct B4/B3/B2 channel order but maps a fixed $[0, 2.8]$ range into RGB. The resulting inputs are unusually dark: archived channel means are about 10.6–11.9 on a 0–255 scale and 99th percentiles are only about 21–25. The original prompt refers to satellite scenes without explicitly naming Local Climate Zones. Raw S1 predicts only six of 17 classes, assigns 89.94% of images to one class, and has Macro-F1 0.0108. Supervised logit adaptation improves this to 0.0597 but does not establish useful scene recognition.

Table 25: Development-only input/prompt repair. Selection uses the mean of the five seed-level OOF Macro-F1 scores; the other diagnostics are archived pooled predictions. All candidates fail the fixed Macro-F1 gate of 0.147059.

Rendering / prompt	Mean OOF F1	Class cov.	Max share %	Entropy
Fixed range + bare	0.07110	17	24.86	0.7471
Fixed range + LCZ	0.05750	17	24.08	0.8099
Min-max + bare	0.14136	17	13.10	0.8932
Min-max + LCZ	0.14631	17	14.12	0.8695

The prespecified gate also requires full predicted-class coverage, maximum class share at most 35%, and normalized prediction entropy at least 0.75. The best arm reaches 0.14631, below the required 0.147059. No arm is selected and no repaired So2Sat test is opened. A versioned finalizer erratum corrects the selection statistic from probability-ensemble F1 to the prescribed mean of five seed-level F1 values; both versions lead to HOLD. We retain this failure instead of relaxing the threshold.

The threshold was defined using a heuristic multiple of $1/17$; this is not a distribution-free expected Macro-F1 for every random predictor. Likewise, a keyword diagnostic flags 48% of generated rules as containing potentially irrelevant content, but it is not a human semantic-accuracy measurement. The original confirmation remains a low-base-performance stress test.

K EXACT RULE-GENERATION AND SCORING SPECIFICATION

K.1 MODEL IDENTITY AND DETERMINISTIC INPUTS

The generator and scorer are the frozen Qwen2.5-VL-7B-Instruct model. The archived configuration hash begins `77d9ec7321cc572e` and the safetensors-index hash begins `73b333b0b16e5286`; full SHA256 values are in `supplement/source_data/settings.json`. These hashes identify archived configuration artifacts; an index hash alone is not a replacement for verifying every model-weight file.

On the rule-generation partition, each source class selects its most frequent wrong predicted class as target. If there is no such confusion, the fallback competitor is selected by aggregate class evidence. Ties follow deterministic native-class ordering. The source and target groups each contribute three images, ranked by own-class evidence with sample ID as a tie-break. The chat input lists SOURCE GROUP (Group A), its three images, CONFUSED GROUP (Group B), its three images, and the generation instruction. The following ordered attempts are part of the frozen protocol; they are not alternative prompts selected by test performance.

Generation attempt 1.

```
Return JSON only. Give up to four rules with one or two short visible-property phrases that
    occur in SOURCE GROUP and distinguish it from CONFUSED GROUP. Never use category or
    group names. Schema: {"rules":[{"clauses":["observable visual property"]}]}
```

Generation attempt 2.

```
Inspect the images and return one to four conservative rules. Each clause must state directly
    observable color, texture, shape, arrangement, or object evidence. Do not state category
    names, functions, locations, or causes. JSON only using the requested rules/clauses
    schema.
```

Generation attempt 3.

```
Return JSON only with at least one conservative visual rule for SOURCE GROUP versus CONFUSED
    GROUP. Use one or two short observable clauses and never mention either class name.
```

Generation uses `do_sample=False` and `max_new_tokens=700`. It does not sample at a selected temperature. The parser enforces the requested structure and clause limits. Syntactic acceptance does not imply a correct visual distinction, as the complete catalog demonstrates.

K.2 BINARY SCORING PROMPTS

EuroSAT and So2Sat System-1.

```
Is this satellite image best described as {class_name}? Answer only Yes or No.
```

Separately frozen AID System-1.

```
This is a satellite land-use image. Is its correct class '{class_name}'? Answer only Yes or
    No.
```

Visible clause scoring.

```
Does this satellite image visibly show the following property? {clause} Answer only Yes or No.
```

Class evidence and clause evidence use the frozen Yes/No scoring interface. There is no extra question mark inserted after the clause placeholder. For multi-clause rules, the minimum clause score is the rule score; the maximum eligible rule score supplies class-rule evidence. Proposal ties are resolved deterministically by the archived strength/class/rule-ID ordering. Numeric code fixes are retained as versioned errata rather than silent changes to frozen outputs.

K.3 TRAINING AND EVALUATION SEQUENCE

Numerical adaptation and feature construction. Binary class evidence q_c is clipped to $[10^{-6}, 1 - 10^{-6}]$ and converted to $z_c = \log(q_c/(1 - q_c))$. The implemented adapted logits are $z_c/T + b_c$,

not $(z_c + b_c)/T$. The L-BFGS-B objective is development cross-entropy plus $10^{-4}\,\mathrm{mean}(b_c^2)$, with $T \in [0.25, 4]$. Optimization coordinates for the class bias have bounds $[-5, 5]$ and are mean-centered to enforce zero sum. Router features contain three scalars (confidence, margin, rule strength), C class-rule scores, two C-dimensional class one-hot vectors, and the changed-proposal flag: $3C + 4$ entries. Logistic heads use $C = 1$, balanced class weights, L1 regularization, the liblinear solver, and at most 3,000 iterations.

1. Freeze data IDs, native class order, rendering, prompts, model identity, and scoring settings; audit authorized partition overlap.
2. Use only the dedicated rule-generation images to choose confusion edges, representatives, and clauses. Save the generated text and parser outcomes.
3. Score development images, cross-fit supervised logit adaptation, and construct deterministic rule proposals. Form benefit/harm targets from development labels.
4. Fit the five router replicas and select the cutoff using the frozen development objective and coverage grid. Freeze the full fitted state.
5. Score the authorized confirmation set once; apply the frozen decision rule without using confirmation labels for selection.
6. Compute paired image-level metrics and retain positive, negative, and HOLD results. Additional ablations remain labeled as such.

L Complete frozen V2 natural-language rule catalog

These tables reproduce every frozen EuroSAT/So2Sat V2 rule in the source catalog, not a curated set of successful examples. Source is the proposed class and target is the confused class determining eligibility. Clauses are preserved verbatim. Some are broad, weak, or physically questionable; their presence is a limitation of the generator, not a reason to rewrite the experimental record after evaluation. The machine-readable CSV also retains original rule IDs and ordering.

Table 26: Complete frozen EuroSAT rule catalog. Source is the proposed class; target is the class that triggers eligibility. Text is reproduced verbatim, including visibly weak clauses. IDs and exact clauses are supplied in the source CSV.

Source	Target	Clause(s)
AnnualCrop	SeaLake	grid pattern
AnnualCrop	SeaLake	square shapes
AnnualCrop	SeaLake	text on image
AnnualCrop	SeaLake	green color
Forest	SeaLake	green color
Forest	SeaLake	mountains
Forest	SeaLake	trees
HerbaceousVegetation	Forest	visible texture
Highway	Industrial	no parking lots
Highway	Industrial	no buildings
Highway	Industrial	no factories
Highway	Industrial	no roads
Industrial	Residential	buildings are large
Industrial	Residential	buildings are close together
Industrial	Residential	buildings are far apart
Industrial	Residential	buildings are small
Pasture	Forest	fields are visible
Pasture	Forest	houses are visible
PermanentCrop	Industrial	fields
Residential	Forest	water
Residential	Forest	roads
Residential	Forest	trees
Residential	Forest	buildings
River	SeaLake	a road is visible
River	SeaLake	a city is visible

EuroSAT rule catalog (continued)

Source	Target	Clause(s)
River	SeaLake	a bridge is visible
SeaLake	Pasture	blue background
SeaLake	Pasture	no trees
SeaLake	Pasture	no buildings
SeaLake	Pasture	text on image

Table 27: Complete frozen So2Sat rule catalog. Source is the proposed class; target is the class that triggers eligibility. Text is reproduced verbatim, including visibly weak clauses. IDs and exact clauses are supplied in the source CSV.

Source	Target	Clause(s)
Compact high-rise	Sparsely built	the letter 'a' is present
Compact high-rise	Sparsely built	the letter 'b' is present
Compact high-rise	Sparsely built	the letter 'c' is present
Compact mid-rise	Sparsely built	the black square is on the left side of the image
Compact mid-rise	Sparsely built	the white square is below the black square
Compact mid-rise	Sparsely built	the black square is above the white square
Compact mid-rise	Sparsely built	the white square is on the right side of the image
Compact low-rise	Sparsely built	the image contains a black background AND the image contains a white square
Open high-rise	Sparsely built	the square is black
Open high-rise	Sparsely built	the square has a white border
Open high-rise	Sparsely built	the square is small
Open mid-rise	Sparsely built	the letter 'a' is present
Open mid-rise	Sparsely built	the letter 'c' is present
Open mid-rise	Sparsely built	the letter 'b' is present
Open low-rise	Sparsely built	the letter 'a' is present AND the letter 'b' is present
Open low-rise	Sparsely built	the letter 'c' is present AND the letter 'd' is present
Lightweight low-rise	Sparsely built	the letter 'a' is present AND the letter 'b' is present
Lightweight low-rise	Sparsely built	the letter 'c' is present AND the letter 'd' is present
Large low-rise	Sparsely built	the image contains a mouse
Large low-rise	Sparsely built	the image contains a keyboard
Large low-rise	Sparsely built	the image contains a computer monitor
Sparsely built	Dense trees	the image is dark
Sparsely built	Dense trees	the image contains a white object
Sparsely built	Dense trees	the image has a black background
Heavy industry	Sparsely built	the white square has a black border
Heavy industry	Sparsely built	the black square is on the left side of the image
Heavy industry	Sparsely built	the black square has a white border
Heavy industry	Sparsely built	the white square is on the right side of the image
Dense trees	Sparsely built	the letter 'a' is present AND the letter 'b' is present
Dense trees	Sparsely built	the letter 'c' is present AND the letter 'd' is present
Scattered trees	Sparsely built	the image is dark
Scattered trees	Sparsely built	the image contains a white object
Scattered trees	Sparsely built	the image has a black background
Scattered trees	Sparsely built	the image is blurry
Bush or scrub	Sparsely built	the image is dark
Bush or scrub	Sparsely built	the image is blurry
Bush or scrub	Sparsely built	the image is clear
Bush or scrub	Sparsely built	the image is bright
Low plants	Sparsely built	the image is dark
Low plants	Sparsely built	the image contains a white object
Low plants	Sparsely built	the image has a black background
Bare rock or paved	Sparsely built	the text is clear
Bare rock or paved	Sparsely built	the text is white
Bare rock or paved	Sparsely built	the black background is present
Bare soil or sand	Sparsely built	the image has a black background
Bare soil or sand	Sparsely built	the image is dark
Bare soil or sand	Sparsely built	the image contains a white object

So2Sat rule catalog (continued)

Source	Target	Clause(s)
Water	Sparsely built	the image is small
Water	Sparsely built	the image is dark
Water	Sparsely built	the image is blurry

M IMAGE-LINKED FROZEN INTERVENTION TRACES

Figure 11 links original evaluated images to their cached clauses and frozen decisions. We choose the lexicographically first sample ID within four outcome strata: EuroSAT active correction and active harm, and So2Sat active correction and active wrong-to-wrong change. So2Sat has no observed active harm, so no such example is invented. Selection is deterministic but outcome-conditioned and retrospective; the examples are not a representative sample or a human-rating study.



a EuroSAT: Correction

Truth: AnnualCrop
Adapted S1: HerbaceousVegetation → AnnualCrop
Clause: "green color"
Clause strength: 0.9627; gate active
Rule ID suffix: db85e088380d3c11



b EuroSAT: Harm

Truth: Forest
Adapted S1: Forest → Pasture
Clause: "fields are visible"
Clause strength: 0.9149; gate active
Rule ID suffix: 4589c4001efa6cdd



c So2Sat: Correction

Truth: Scattered trees
Adapted S1: Dense trees → Scattered trees
Clause: "the image is dark"
Clause strength: 0.9876; gate active
Rule ID suffix: 1726b409229aa856



d So2Sat: Wrong-to-wrong

Truth: Bare soil or sand
Adapted S1: Dense trees → Scattered trees
Clause: "the image is dark"
Clause strength: 0.9876; gate active
Rule ID suffix: 1726b409229aa856

Figure 11: **An explicit trace can accompany either a correction or a failure.** Original RGB pixels are displayed without cropping, contrast adjustment, or image generation (nearest-neighbor enlargement only; EuroSAT 64×64, So2Sat 32×32). Clauses and scores are reproduced from frozen caches. (a) A correction to AnnualCrop. (b) A correct Forest prediction is harmed. (c,d) The same "the image is dark" clause and score accompany a So2Sat correction and a wrong-to-wrong change. Clause strength is the VLM's cached Yes/No score, not a verified probability of semantic correctness.

The archived identifiers are `AnnualCrop_1865`, `Forest_1028`, `so2sat:testing:10038`, and `so2sat:testing:10017`. Full sample/rule IDs, input SHA256 hashes, clause scores, reference predictions, final predictions, and selection criteria are retained in `trace_cases_v2.json`. The proposal is reconstructed by maximizing the cached minimum-clause strength, then negative

source-class ID, then frozen rule ID. Every reconstructed proposal equals the archived final prediction for these active examples; no VLM or router was rerun.

Rule eligibility follows raw S1, not adapted S1. In (a), raw S1 is SeaLake while adapted S1 is HerbaceousVegetation; the AnnualCrop-versus-SeaLake rule supplies the candidate. In (c,d), raw S1 is Sparsely built while adapted S1 is Dense trees. This distinction is necessary to reconstruct the exact computation and prevents interpreting the rule's target class as necessarily the reported pre-intervention class. The records expose the implemented decision path, not a causal explanation of VLM reasoning.

N STATISTICAL INTERPRETATION AND REPRODUCIBLE FIGURE DATA

Sampling unit and pairing. The unit for confirmation intervals is the unique image. Each bootstrap replicate samples image indices with replacement within native ground-truth classes and applies the same sampled indices to both methods. The reported interval is the 2.5th–97.5th percentile of the resulting Macro-F1 differences. There are 10,000 replicates. Router replicas and repeated clause queries are not independent sample units and do not multiply the test sample size.

Seeds and distinct analyses. Primary EuroSAT/So2Sat bootstrap analysis uses seed 20270905; AID final-confirm uses 20270909. Semantic controls use 20270907 and visual controls 20270908. Thus a repeated full-model interval in an auxiliary table can differ slightly from the primary interval while predictions remain identical. The main result uses the original primary interval, not whichever rerun has a more favorable endpoint.

Conditional risk and zero events. Activated-sample harm is $H_A/|A|$, with both numerator and denominator restricted to the same active set and comparator. Overall harm is H/n. Neither is interchangeable with the other. Zero observed harms on a finite test are not a zero-risk certificate. We retain the failed strict-safety history and make no distribution-free risk guarantee. Geographic correlation can also make an image bootstrap optimistic for deployment on a new region; duplicate auditing is not proof of geographic independence.

Post-hoc analyses and protocol evolution. The main V2 operating points precede their confirmation access. The component, semantic, and visual-control analyses were added later on immutable outputs or images. They can test weaknesses in an existing claim, but do not make the entire research process prospectively preregistered. Historical V1, AID final-confirm, and the So2Sat development repair have different roles and are never merged as independent replications of one fixed experiment.

Artifact layout. The accompanying manuscript source contains `supplement/source_data/` with figure/table CSVs and a provenance manifest, `supplement/tables/` with deterministic LaTeX outputs, and `supplement/figures/` with vector PDFs and editable SVG panels. The `build_data.py`, `build_tables.py`, and `build_figures.py` scripts preserve the mapping from locked experimental outputs to displayed numbers. Experimental locks are read-only inputs, not manuscript editing targets. Regeneration of figures does not require new VLM inference.